Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster

Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*

Rebekah L. Rogers and Daniel L. Hartl

Research Article

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Running head: Rapid evolution via chimeric genes

Key words: chimeric genes, *Drosophila melanogaster*, regulatory evolution, evolutionary novelty, adaptive evolution, duplicate genes

Corresponding author: Rebekah L. Rogers, Dept. of Ecology and Evolutionary Biology, 5323 McGaugh Hall,University of California, Irvine, CA 92697

Current affiliations: Rebekah L. Rogers, Dept. of Ecology and Evolutionary Biology, 5323 McGaugh Hall,University of California, Irvine, CA 92697.

Daniel L. Hartl, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Phone: 949-824-0614
Fax: 949-824-2181
Email: rogersrl@uci.edu

Abstract

Chimeric genes form through the combination of portions of existing coding sequences to create a new open reading frame. These new genes can create novel protein structures that are likely to serve as a strong source of novelty upon which selection can act. We have identified 14 chimeric genes that formed through DNA-level mutations in *Drosophila melanogaster*, and investigate expression profiles, domain structures, and population genetics for each of these genes to examine their potential to effect adaptive evolution. We find that chimeric gene formation commonly produces mid-domain breaks and unites portions of wholly unrelated peptides, creating novel protein structures that are entirely distinct from other constructs in the genome. These new genes are often involved in selective sweeps. We further find a disparity between chimeric genes that have recently formed and swept to fixation vs. chimeric genes that have been preserved over long periods of time, suggesting that preservation and adaptation are distinct processes. Finally, we demonstrate that chimeric gene formation can produce qualitative expression changes that are difficult to mimic through duplicate gene formation, and that extremely young chimeric genes ($d_S < 0.03$ are more likely to be associated with selective sweeps than duplicate genes of the same age. Hence, chimeric genes can serve as an exceptional source of genetic novelty that can have a profound influence on adaptive evolution in *D. melanogaster*.

Introduction

Chimeric genes form when portions of two or more coding regions fuse to create a novel open reading frame. Such rearrangements have often been cited as a potential source of novelty upon which selection can act (Patthy 1999; 2003, Rogers, Bedford and Hartl 2009, Gilbert 1978, Peisajovich et al. 2010). Most arguments for the utility of chimeric genes have focused on the effects of exon shuffling with respect to protein structure, with particular emphasis on rearrangement of whole protein domains (Bashton and Chothia 2007, Vogel et al. 2004, Patthy 1999; 2003, Voigt et al. 2002).

Both theoretical modeling and *in vitro* peptide splicing have shown that rearranging protein segments can result in novel peptides that are difficult to reach through alternative means (Giver and Arnold 1998, Cui et al. 2002). By effecting 'leaps' through functional space, chimeric genes may be able to reach new functional constructs that cannot be accessed through duplication and point mutation. Yet, the limits of protein modularity are not fully understood. Changes in smaller units, below the level of whole protein domains may still be able to reach distinct structures, or alternatively could display strong mutational constraint similar to that of point mutations. If protein structures are sufficiently flexible, chimeric genes which recombine smaller protein subunits may be similarly useful in generating novel functions, and may still serve as a source of adaptive changes.

However, the ability of chimeric gene formation to generate novelty goes far beyond obvious structural changes. One chimeric gene in *D. melanogaster*, *Quetzalcoatl* (*Qtzl*), has produced both structural and gene expression changes through a single complex mutation

and is associated with a strong selective sweep (Rogers et al. 2010). Through the combination of regulatory elements and cellular targeting elements, this new gene has emerged with an expression profile that is distinct from its two parental genes (Rogers et al. 2010). Yet, it is uncertain how often such expression changes occur or how they might influence the role of chimeric genes as a source of phenotypic novelty.

We have identified 14 chimeric genes in *D. melanogaster* that formed from the rearrangement of two or more coding sequences (Rogers, Bedford and Hartl 2009), based on strict criteria to identify a set of genes that were clearly of chimeric origin. All of these chimeric genes appear to have formed through tandem duplications that did not respect gene boundaries (Rogers, Bedford and Hartl 2009), in stark contrast with many previously identified chimeric retrogenes in other *Drosophila* species.

Previous studies in other species of *Drosophila* and in humans have shown that many chimeric retrogenes show sequence signatures that are consistent with natural selection acting to fix and modify newly formed chimeras (Shih and Jones 2008, Jones and Begun 2005, Long and Langley 1993, Ohshima and Igarashi 2010). In contrast, chimeric genes identified in *C. elegans* display signals that are consistent with neutral evolution (Katju, Farslow and Bergthorsson 2009). Thus, it is not yet clear whether such sweeps are typical or whether chimeric genes are generally likely to contribute to adaptation. While many new genes in *D. melanogaster* appear to be chimeric (Rogers, Bedford and Hartl 2009, Zhou et al. 2008), it is unknown whether the DNA-level mutations observed in *D. melanogaster* can provide a source of novelty that is comparable to that of chimeric retrogenes or whether they make

similar contributions to genomic content in comparison to duplicate genes.

In order to assess the propensity of chimeric gene formation to produce novel genetic constructs that can be useful in adaptive evolution, we have characterized protein domain structures, mRNA profiles and population genetic parameters for chimeric genes. As a comparison, we provide a similar analysis of duplicate genes and conclude that chimeras are likely to serve as a more immediate source of genetic novelty for adaptation in the short-term.

Methods

Methods for chimera identification

As described previously (Rogers, Bedford and Hartl 2009), we performed an all-by-all BLASTn comparison (Altschul et al. 1990) considering only non-self matches with $E < 10^{-10}$ for the *D. melanogaster* r.5.2-all-CDS data set obtained from FlyBase (accessed August 2007; ftp://ftp.flybase.net/releases/) (Adams et al 2000). Chimeric genes are defined for present purposes as those that fit the following criteria: (a) the two most significant matches identify putative parental genes; (b) one parental gene provides the exons that contribute to the $5'$ end of the candidate chimera and the second parental gene contributes to the remainder of the candidate chimera; (c) the two parental genes must hit regions of the chimera that do not overlap by more than 15 bp; and (d) the chimeric gene must be the best hit for each parent. We removed any genes that physically overlapped with their two parental genes, and we excluded any heterochromatic sequences where assembly and annotations are still

not firmly established. Prior to further analysis, we confirmed the existence of each chimeric gene with PCR amplification from genomic DNA of the *D. melanogaster* reference strain $y^1\ cn^1\ bw^1\ sp^1$. These qualifications produced a final list of 14 putative chimeric genes in *D. melanogaster*.

These requirements are quite strict and are not designed to capture all chimeric genes within the genome. They do not include chimeras derived from transposable elements, formerly non-coding sequences, or transcripts that have recruited novel UTRs but do not contain novel coding sequences. Rather, they define a conservative list of genes whose chimeric origins are relatively certain. None of these chimeric genes appear to be derived from retrogenes, although our definitions did not explicitly exclude such constructs.

The genomic sequence for each *D. melanogaster* chimeric and parental gene was obtained from FlyBase and aligned using a blast2seq (Tatusova and Madden 1999) to determine the breakpoints of chimera formation. Genomic sequences of parental genes were also aligned to one another, although no significant similarity was found at the nucleotide level. We aligned the translation of each chimeric gene with its parental sequences, back-translated to produce in frame alignments and calculated d_S using PAML. We then used a Bayesian framework to correct estimates of d_S for the effects of sequence length (Rogers, Bedford and Hartl 2009). Seven of these chimeric genes have $d_S < 0.03$ and are all unique to *D. melanogaster*, suggesting that they are exceptionally young. The remaining seven genes are classified as 'older'; all have orthologs in at least one other *Drosophila* species and have $d_S > 0.1$.

We identified duplicate genes in a similar manner. Duplicate genes were defined as

reciprocal best hits that did not overlap physically. We removed all genes that match to heterochromatic sequences and we excluded duplicates from large gene families of more than five members. Translations were aligned for each paralog, then back-translated to produce in frame alignments that were used to calculate d_N and d_S in PAML (Rogers, Bedford and Hartl 2009). Estimates of d_S were then corrected for effects of sequence length (Rogers, Bedford and Hartl 2009). These requirements provided a list of 37 young duplicate genes with $d_S < 0.03$.

Establishing expression profiles

We obtained expression data based on uniquely mapped paired-end reads from the modENCODE gene expression project for each parental and chimeric gene and for each duplicate gene pair. The modENCODE data include expression profiles obtained via RNA-seq for multiple developmental time points, including adult males and adult females. Additionally, we performed RNA extractions and RT-PCRs on heads, testes plus accessory glands, and carcass of adult males.

We extracted RNA using a standard phenol-chloroform protocol. We pulverized whole or dissected flies in 1 mL of TRIzol (Invitrogen) and incubated at room temperature for 5 min For whole flies or carcasses, the TRIzol suspensions were centrifuged 10 min to precipitate exoskeleton material. TRIzol suspension was added to 200 μL of chloroform in a phase-lock gel tube (Eppendorf) and agitated for 30 s, then left at room temperature for 3 min. Samples were centrifuged for 10 min at 4℃. The top, clear phase was removed and added to 500

μL of isopropanol and mixed by inversion for 10 min, then centrifuged for 10 min at 4℃. Supernatant was removed, and the pellet was washed with isopropanol and centrifuged again for 10 min at 4℃. Isopropanol was removed, and the pellet was washed with 75% ethanol and centrifuged for 10 min at 4℃. Ethanol was removed and the pellet was allowed to dry. Nucleic acid was suspended in nuclease-free water.

We diluted RNA to a concentration of approximately 20 ng/μL, treated with Turbo DNase (Ambion) according to the standard protocol, and prepared cDNA from 7 μL of DNase-treated RNA using an oligo dT 18mer with the SuperScript II system (Promega). The cDNA preps were then amplified via PCR, using gene specific primers. PCRs were as follows: 95℃ 5 min, followed by 40 cycles of 95℃ for 40s, 45℃ for 45s, 72℃ for 60s, ending with a final extension of 72℃ for 5 min Four genes, *CG18853*, *CG31904*, *CG30457*, and *CG6653* contained low-complexity regions and could not be readily assayed in adult tissues due to difficulties of primer design in duplicate sequences. Combined results of modENCODE and RT-PCRs are available in Table 2 and Tables S1-S13.

Identifying protein domains and target sequences

We used the TargetP 1.1 webserver (http://www.cbs.dtu.dk/services/TargetP/, Oct 22, 2009) (Emanuelsson et al. 2000) to identify mitochondrial and secretory target peptide sequences in the longest translations for each chimeric and parental gene. We identified membrane-bound domains using the HMMTop webserver (http://www.enzim.hu/hmmtop/ May 24, 2010) (Tusnady and Simon 1998; 2001). Results were verified with published

empirical observations when available. We used the Pfam database to identify all known protein domains for each parental and chimeric gene (http://pfam.sanger.ac.uk/). Orthologs of chimeras and parental genes were identified throuh a BLASTp search (Altschul et al. 1990) against the non-redundant protein database
(http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Divergence and polymorphism analysis

Selective sweeps cause a single allele to spread rapidly through a population, reducing nucleotide diversity at the locus of selection. Linked sites near the selected locus will also display decreased diversity, although this signal tends to decay as recombination reduces linkage disequilibrium, leading to a classic V-shaped graph of diversity (Hartl and Clark 2007). Selective sweeps additionally modify the site frequency spectrum, causing an excess of singleton sites relative to moderate frequency sites. Such changes in the site frequency spectrum are reflected in Tajima's D, which measures deviations in the average number of pairwise substitutions relative to the number of segregating sites (Tajima 1989). Reduced diversity and negative values of Tajima's D are commonly used to detect selective sweeps.

We identified all polymorphisms on chromosomes 2 and 3 in the DPGP *D. melanogaster* Solexa Assemblies Release 1.0 of 37 sequenced strains of *D. melanogaster* from Raleigh, North Carolina (accessed Jan 2011; http://www.dpgp.org/). This dataset provides aligned sequence for all uniquely mapping reads in the genomes of 37 strains of *D. melanogaster* from Raleigh. We calculated π, θ_W and Tajima's D (Tajima 1989) for 10 kb windows sliding

along chromosomes at 1 kb intervals, using only sites that had mapped reads for all Raleigh strains. We considered only windows with more than 1 kb of sites with full coverage.

Kaplan et al. (Kaplan, Hudson and Langley 1989) present a series of differential equations that describe the expectation of π for a neutral locus linked to a selected locus that has undergone a recent selective sweep. These equations do not admit closed-form solutions. Using Mathematica v6.0, we solved these equations numerically to find the expectation of π moving outwards from a selective sweep with selective coefficient s and time of fixation t_f. This model assumes a single sweep and simple demography. While the North American population has experienced a bottleneck in the expansion out of Africa, estimates of selection coefficients should be fairly robust, especially for sweeps that occurred prior expansion of *D. melanogaster* into Europe and North America.

We fit these expectations to the observed genetic diversity surrouding *CG18217* on chromosome 3L. We assumed a neutral mutation rate μ of 5.8×10^{-9} substitutions per site per generation (Haag-Liautard et al. 2007), a generation time τ of 10 generations per year, an effective population N_e of 1.85×10^6, based on coalescent analysis (Rogers et al. 2010), and rate of recombination r of 1.32×10^{-8} events per site per generation, obtained from the *Drosophila melanogaster* Recombination Rate Calculator v2.1 (http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl) (Singh, Arndt and Petrov 2005). Background diversity measures were set to the chromosome 3L average of $\pi = 0.006172$.

Similarly, we fit the sweep model to the observed genetic diversity surrounding *CG18853*

on chromosome 2R. The recombination rate in this region was set to $r = 0.90 \times 10^{-8}$ (http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl) (Singh, Arndt and Petrov 2005). *CG18853* lies in a section of chromosome 2R that has lower average diversity than the remainder of the chromosome. Hence, we set background diversity equal to the average diversity from 900 kb to 6,000 kb ($\pi = 0.004348$) rather than the whole chromosome average ($\pi = 0.006336$). For both selective sweeps, parameters s and t_f were estimated through least squares minimization. These estimates assume that regions surrounding the chimeric gene are completely neutral and that demography is simple, i.e. population size has remained constant through time.

A total of three young chimeric genes, *Qtzl* (Rogers et al. 2010), *CG18217* and *CG18853*, are located at the center of selective sweeps out of seven newly-formed chimeras. However, randomly picking seven genes from the autosomal regions might theoretically result in a similar number of selective sweeps. To account for this possibility, we produced 10,000 replicates of seven genes chosen without replacement from the two *D. melanogaster* autosomes. Heterochromatic regions were excluded from the initial chimeric gene search, and therefore these were not included in resampling. The centromeric and telomeric regions show abnormal diversity patterns and site frequency spectra compared to the majority of the autosomes, and hence genes from these locations were also excluded. Tajima's D was recorded for each gene sampled as a measure of selection. A similar approach was used to compare the role of chimeric and duplicate genes in selective sweeps. We produced 10,000 replicates of seven genes chosen at random without replacement from the set of

young duplicate genes which do not lie adjacent to chimeric genes.

These criteria provide an underestimate of the contribution of chimeric genes in adaptation, as our power to detect partial selective sweeps or sweeps from standing variation is extremely limited. Similarly, we are not able to examine the full contribution of the seven ancient chimeric genes, as any sweeps to fixation just after formation occurred so far in the past that standard selective sweep signals will have degraded completely.

Molecular Evolutionary Analysis

We estimated d_N and d_S along the branches of the phylogenetic tree for six of the seven chimeric genes. Current ortholog calls show only a single ortholog for *CG6653*, which was not sufficient to provide accurate estimates of d_N and d_S, and thus it was excluded. Assembly of chimeric genes can be problematic, and it is not entirely certain that these represent true absence of an ortholog in all species. Coding sequences corresponding to orthologs of each chimeric gene were obtained from FlyBase (June 2011). Translations of the chimera and its orthologs were aligned and then back-translated to generate in frame alignments for each chimera ortholog set. We used the codeml package in PAML to estimate d_N and d_S. For each gene we assumed no across-site rate variation (α parameter set equal to infinity), estimated transition-transversion bias from each gene (estimate κ), and calculated equilibrium codon frequencies based on overall nucleotide frequencies (F1×4). In some cases a subset of all reported orthologs were used due to saturation of synonymous sites or ambiguous alignments among the oldest orthologs which might interfere with estimates of substitution rates.

Results

Full length duplicate genes typically carry promoters, UTRs, RNA stability sites, and cellular targeting signals that are identical to their original singleton parents. As such, they will often emerge with an expression profile that is identical or highly similar to that of their ancestral sequences. While dispersed duplicates and retrogenes offer a greater chance to recruit novel regulatory elements, their ability to change within-transcript regulatory elements may still be limited. Chimeric genes, however, combine portions of different transcripts, and as such they can shuffle regulatory elements and emerge with mRNA profiles that are distinct from their parental genes. This ability to generate regulatory novelty in addition to peptide changes can offer a wider range of phenotypic and adaptive outcomes than gene duplication.

We describe expression profiles, cellular targeting, protein domains, and population genetics of the chimeric and parental genes to determine the factors that influence the selective impacts of chimeric genes in *D. melanogaster*.

Expression

We previously identified 14 chimeric genes in *D. melanogaster*. Seven of these genes are exceptionally young ($d_S < 0.03$) and are specific to *D. melanogaster*; the remaining seven are older and appear to have been incorporated stably into the genome (Rogers, Bedford and Hartl 2009). Among the youngest chimeric genes we can determine the expression consequences of chimeric gene formation by comparing the chimera to its parental genes. These young genes are newly formed and have not had sufficient evolutionary time to

accumulate substantial nucleotide changes after formation. All regulatory material is inherited from either the 5′ or 3′ parental gene, and hence chimera expression patterns should reflect those of the parental genes. However, through the shuffling of different promoters, enhancers, and RNA stability sites, it is possible to create a number of distinct expression profiles.

Some chimeric genes display expression patterns that closely resemble only one parental gene. For example, *CG18853* has an expression profile that closely resembles the parental gene that donated the 5′ end and promoter (Table 1). Here, chimeric gene formation resulted in a novel peptide that now appears in parallel with one parental gene, and the 3′ parental gene contributes little to expression patterns. Similarly, *CG32318* takes a portion of the 3′ parental gene, *CG9187*, an S phase regulator, and places it in an expression profile that mimics *CG9191*, a kinesin protein involved in microtubule movements (Table 1). The peptide is placed in a more limited context than its 3′ parent, thus allowing for specialization. The change appears to be neutral in this particular case, but the general phenomenon could have broad impacts on pleiotropic and selective constraints.

In some cases, however, chimeric genes can create fully distinct expression profiles through the shuffling of regulatory elements in the 5′ and 3′ ends. The young chimeric gene *CG12592* pulls the majority of its genetic material from the parental gene *CG12819*. Across tissues, it is expressed similarly to the parental gene that formed its 5′ end, *CG18545*, but its expression pattern across developmental time points is identical to that of the gene that donated the 3′ segment, *CG12819* or *sle*, a peptide necessary for brain development (Table 2).

Here, we find that the peptide sequence belonging to *sle* has changed context across tissues while maintaining its expression profile across time points. Hence, enhancers or stability sites which govern expression in male tissues must act independently of sites governing expression during development. Additionally, *Qtzl*, a chimeric gene that was involved in a recent selective sweep, has an expression profile that is distinct from either parental gene. The expression profile largely mimics that of the 3′ parent, *escl*, but due to the novel combination of regulatory elements, it is expressed in the heads of adult males as well as in late embryonic stages (Rogers et al. 2010; Table S4).

Finally, one chimeric gene is expressed in cases where both parental genes are silenced. *CG11961* shows expression in the testes, late larvae, and whole adult females, in contrast to both parentals. It is expressed in every tissue and life stage examined, with the exception of newly fertilized embryos less than two hours old (Table S5). Assuming that this chimeric gene formed through tandem duplication (Rogers, Bedford and Hartl 2009), the gene that donated the 3′ end of the gene has relocated, giving *CG11961* new upstream material and *CG30049* new downstream material. At present *CG30049* is expressed only in pupae and in the carcass of adult males (Table S5). *CG9416*, which donated the 5′ end of the gene is expressed across most stages and tissues, but is not found in testes, late larvae, or whole adult females (Table S5). Thus, *CG11961* has an expression profile that is unique from its parental genes on several points. Whether this expression represents neofunctionalization or partitioning of ancestral expression patterns cannot be determined from present data.

As a comparison for these newly formed chimeras, we used the modENCODE unique

mapping data set to identify mRNA profiles for our 37 duplicate gene pairs with $d_S < 0.03$. Seven of these did not have modENCODE expression data because of changes in gene annotations, possible gene silencing, or lack of uniquely mapping reads. From the remaining 30, only one duplicate gene pair shows evidence for qualitative differential expression comparable to that observed in chimeric genes, although even this change involves a relatively minor difference in the timing of expression (Table S14). A Fisher's exact test of these ratios yields $P = 0.0388$. Hence, while duplicate genes may be able to produce quantitative changes in gene expression, their ability to generate novel expression profiles is extremely limited in comparison to chimeric genes. Thus, chimeric genes may be a richer source of genetic novelty that can influence evolutionary outcomes in profound ways.

Targeting

Beyond changes in transcription and RNA stability, different peptides can be targeted to different compartments within the cell, opening up a greater diversity of profiles and functions. For example, *Qtzl* inherits a mitochondrial target sequence from the 5′ parental gene, allowing this new sequence to be targeted to the organelle (Rogers et al. 2010). Based on TargetP predictions, several other chimeric genes have experienced targeting changes.

The chimeric gene *CG31687* appears to inherit a mitochondrial target signal from the 5′ gene *CG2508*, while its 3′ parental *CG31688* is targeted to the cytoplasm. Conversely, the chimeric gene *CG18217* appears to be targeted to the cytoplasm whereas the 3′ parent *CG4098* is predicted as a secreted peptide, a signal that likely reflects a nuclear targeting

signal for the *CG4098* DNA repair peptide. The 5′ parent of *CG18217*, designated *CG17286* is also targeted to the cytoplasm. These changes can broaden or narrow the cellular context of a particular peptide, influencing phenotypic outcomes as well as subsequent evolution.

Some of these chimeric genes appear to be selectively favored, whereas others are consistent with neutral processes. Regardless of the selective impacts of individual genes, the ability of chimeric genes to modify the cellular context of a peptide and to target a sequence to various cell compartments should allow for a diverse range of phenotypes as a consequence of mutations. As typical duplicate genes carry the entire protein sequence of their singleton ancestors, they will be unable to effect similar changes in cellular targeting. Hence chimeric gene formation should be able to affect a wider range of phenotypic outcomes than gene duplication.

Shuffling membrane-bound domains

Classic views on exon-shuffling have focused largely on recombination of whole conserved protein domains. However, changes on a finer scale below functional domains may be equally important in the development of novel peptide structures. Membrane-bound domains provide short, modular units whose presence or absence can significantly impact peptide functions (Tusnady and Simon 1998). We used the HMMTop webserver to identify membrane bound domains in each of our chimeric and parental genes to explore the potential for changes in membrane anchoring and orientation.

CG31904 contains a major part of *Acp1*, but with orientation reversed. *CG13796*

is predicted to have neurotransporter activity and has a total of twelve predicted transmembrane helices. Adult cuticular protein 1 (Acp1) is a cuticular protein component expressed in the heads and thorax or adult *D. melanogaster* (Qiu and Hardin 1995). The chimera inherits three transmembrane helices from the neurotransmittor and one transmembrane domain from *Acp1*. All three are predicted to have an N-terminus inside the cell. The resulting protein carries the majority of *Acp1* now oriented inside, rather than outside the cellular membrane (Fig. 1).

Based on a worldwide sample of *D. melanogaster*, the chimeric gene appears to be absent in many lines, and measures of nucleotide diversity and site frequency spectra suggest that these particular changes were likely neutral, or nearly neutral, consistent with the general inert properties of cuticular peptides. We see no evidence of pseudogenization that might suggest ancient origins. Yet again, this particular type of change where portions of proteins change transmembrane status and orientation through the combination of different proteins could in some cases produce structures with unique functional attributes, especially when modifying more biochemically active proteins.

Mid-domain breaks

Similarly, much of the exon-shuffling literature asserts that recombination between domains is far more likely to be favorable than mid-domain breaks. However, recent work has shown that breakpoints within domains can produce functional peptides as well (Mody, Weiner and Ramanathan 2009). We examined chimeric genes to assess their propensity to generate and

tolerate breakpoints within conserved protein domains rather than whole domain shuffling.

Of the 7 youngest chimeric genes, we have found three where breakpoints occur within, rather than outside of protein domains. *CG31904* disrupts a sodium neurotransmitter domain and pairs this segment with *Acp1*. As discussed above, this change has altered the membrane orientation of the protein, reflecting protein modularity in secondary structure below the domain level. Similarly *CG18853* displays a mid-domain break that disrupts an uncharacterized conserved domain as well as an FAD-binding segment (Fig. 2). Finally, *CG32318* breaks apart a kinesin domain, combining this $5'$ end with a portion of a cell regulatory peptide. All seven young chimeric genes form from unrelated peptide sequences which house fully distinct domains. These results suggest that the functional units of peptide modularity lie at a smaller scale than previously thought and that peptide structures may be fairly amenable to modification.

When we examine the older chimeric genes that are stably incorporated into the genome, we see a very different pattern. All of the preserved chimeras where protein domain data is available in Pfam have formed from parental proteins that display amino acid similarity to the same conserved domains and all align in a BLASTp. Many of these chimeras and parental genes have distinct expression profiles that prevent total functional overlap. However, conserved domains appear to be identical, and the locations of membrane-bound helices predicted by HMMTop appear to be the same. Using a Bayesian binomial approach, we can determine that the one-sided 95% CI for the rate at which chimeras form from related parental genes must lie below $p = 0.26242$. The probability of choosing 7 of these

to be retained and choosing none of the other types is $P < 0.26242^7 < 10^{-5}$. Hence, the overrepresentation of chimeras formed from similar peptides among the preserved chimeras is extremely significant.

All preserved chimeric genes where d_N and d_S estimates are available show strong constraint in amino acid substitutions on multiple branches of the tree, suggesting that chimeras are not entirely functionally redundant with their parental sequences. One chimera, *CG31688* may show signs of higher substitution rates on the most recent branches, although alignment is ambiguous for short sections of the gene in *D. simulans* and *D. sechellia* possibly inflating d_S. The gene still displays constraint on all ancestral branches.

The methods we used to identify chimeric genes are biased against new genes that subsequently duplicate to form their own gene families. As such, adaptive chimeric genes which have proliferated within the genome may be underrepresented, partially explaining the discordance. We modified our search of chimeric genes to allow for subsequent duplication, and found that all of the older chimeric genes still form from highly similar parents (SI Text). Hence, the disparity is not due to this particular aspect of our chimeric gene identification methods.

Selection

If the formation of chimeric genes is indeed a key contributor to adaptive evolution, then we should observe signals of positive selection surrounding the youngest chimeric genes. After selective sweeps, where a favored sequence spreads quickly through the population, we should

observe statistical signals that include reduced nucleotide diversity and highly skewed site frequency spectra (Tajima 1989).

The chimeric gene *CG18217* appears to have formed recently in *D. melanogaster* and is not shared with any other *Drosophila* species. In spite of having formed very recently, it appears to have risen to high frequency worldwide. It is found in 9/10 African strains, and 11/12 strains from a worldwide collection. Furthermore, 37/37 Raleigh strains from the DPGP release 1.0 show sequencing reads that span the unique chimera boundary. Assuming that presence or absence in each strain is an independent Bernoulli trial, and given a uniform prior distribution on population frequency, we estimate the frequency *CG18217* worldwide falls between 0.8847 and 0.9896 (95% Two-sided CI).

CG18217 also lies near the bottom of a wide valley in diversity on chromosome 3L (Fig. 3). Tajima's D in the region approaches -2.5, indicating highly skewed site frequency spectra. The reduction in diversity spans roughly 40 kb, and the chimeric gene lies at the center of the sweep. The lower boundary is abnormally flat and wide, with sharp slopes, a product of the low recombination rate within this region (Singh, Arndt and Petrov 2005). Fitting the Kaplan-Langley equations, which assume a single sweep and simple demography, the sweep appears to have occurred around 20,000 years ago just prior to the migration out of Africa with a selective coefficient of 0.6%.

CG18217 has been changed to a pseudogene annotation in the most recent *D. melanogaster* genome releases. Yet, the gene clearly aligns to known ESTs and transcripts have been amplified using polyA preparations and show the presence of correctly spliced

introns. Furthermore, the associated coding sequence contains no premature stop codons. Considering the current evidence, as well as its presence in a region which has experienced a selective sweep, we expect that this gene could well be functional. The 5′ end of *CG18217* is derived from *spd-2*, an essential component of the centrioles required for formation of the spindle. It is active during the earliest stages of mitosis and meiosis (Giansanti et al. 2008). The *spd-2* mRNA is strongly expressed in developing embryos, pupae, and adult females. It appears to be moderately expressed in whole adult males (Table S1). *CG18217* contains a NUDIX DNA repair domain in its 3′ end, which is derived from the parental gene *CG4098* (Fig. 4). *CG4098* is most strongly expressed in pupae and adult females (Table S1). We were unable to amplify *CG4098* from cDNA derived from adult male testes or carcasses, although it was successfully amplified from cDNA from adult heads (Table S1). *P*-element insertions for *CG18217* are listed in FlyBase as viable and fertile, as would be expected for a newly-formed gene with partial redundancy in the genome.

It is entirely possible that the combination of this DNA-repair domain with a regulatory element that functions just before cell division could be advantageous in preventing cellular errors. NUDIX domains have also been implicated in small molecule signaling (McLennan 1999), which could produce new phenotypic effects. Alternatively, epistatic interactions between the separate sections of the peptide, a common consequence of domain tethering (Bashton and Chothia 2007), could result in a new function.

Another chimeric gene, *CG18853*, also lies in a valley of reduced diversity and shows skewed site frequency spectra on chromosome 2R. The reduction in diversity appears to

span roughly 45 kb (Fig. 5). Such a reduction in diversity is consistent with a single sweep occurring around 200,000 years ago with a selection coefficient of 0.25%. *CG18853* houses portions of two protein domains but is characterized by an unusual breakpoint that lies within two domains. Whether or not this gene is functional remains uncertain. Transposable element insertion lines are listed in FlyBase as viable and fertile, however, again, this is expected for a newly formed gene.

The parental peptide CG12822 carries a conserved domain of unknown function that is found in vertebrates as well as in multiple bacteria. The human ortholog of *CG12822*, Nef-associated protein 1, is a thioesterase that interacts with HIV protein Nef (Liu et al. 1997). The remainder of the peptide contains an FAD-binding domain derived from a photolyase (*phr*). The boundary of chimera formation falls within these two domains, resulting in a chimera that combines portions of domains rather than whole domain shuffling (Fig. 2). How the different portions of these domains interact is not known. Still, resistance to viruses and similar pathogens could create an opportunity for an evolutionary arms race which might generate strong selection to fix new genes but result in selective pressures that are transient, consistent with the patterns observed in chimeric genes.

In each of these cases we cannot be entirely certain that the locus of selection lies in the chimeric gene, a common problem with scans of selection. Furthermore, recent fixation of tightly-linked duplicate genes through neutral processes can cause moderate reductions in diversity and Tajima's D (Thornton 2007). These types of effects may explain a portion of the signals that we see, but would be insufficient to produce reductions of this breadth or

magnitude.

Two other chimeras, *CG12592* and *CG31668*, display a less drastic reduction in diversity and somewhat skewed site frequency spectra (Table 3) but both are several kb away from the local minimum, and are not strong candidates for selective sweeps.

Qtzl, *CG18217*, and *CG18853* are all newly-formed chimeric genes that are found at the center of selective sweeps in *D. melanogaster*. We used a resampling approach, choosing seven genes at random from the two *D. melanogaster* autosomes, to account for the likelihood of finding three selective sweeps among seven genes. Using a fairly liberal cutoff of Tajima's $D < -1.8$, we found 39 out of 10,000 replicates had three or more genes that might potentially be involved in selective sweeps. This cutoff is far less stringent than that applied to any of our chimeric genes. It does not require that the gene be an outlier with respect to the region around it, and few of these supposed sweeps have the same breadth as those found at our chimeric genes. Thus, the likelihood of obtaining similar results by chance must be exceedingly rare, and is most certainly $P < 0.0039$.

Comparing each of these chimeric genes to their parental sequences, we find that they each have $d_S < 0.03$. During this time frame, an estimated 15.5 chimeric genes will have formed (Rogers, Bedford and Hartl 2009), suggesting that 19.3% of chimeric genes are subject to selective sweeps just after formation. This contrasts with a frequency of preservation of 1.4% (Rogers, Bedford and Hartl 2009). This disparity between the frequency of fixation due to selective sweeps and the frequency of preservation, combined with the disparity in domain structures for newly formed and preserved chimeric genes, strongly suggests that

adaptation and gene preservation are largely distinct phenomena (see Discussion).

In contrast, out of 37 pairs of duplicate genes with $d_S < 0.03$ not located in regions where chimeric genes formed we found 4 pairs with Tajima's $D < -1.8$. In this time frame an estimated 104.1 duplicate genes will have formed (Rogers, Bedford and Hartl 2009), suggesting that the frequency with which new duplicate genes are involved in selective sweeps is only 3.8%. Again, this requirement is far less stringent than the criteria used for selective sweeps on chimeric genes and may overestimate the contribution of young duplicate genes in adaptation. We performed 10,000 jacknife replicates, choosing 7 genes without replacement from the list of duplicate genes with $d_S < 0.03$. In 10,000 jacknife replicates, 168 had 3 or more pairs with Tajima's $D < -1.8$, indicating that $P < 0.0168$. Hence, the overrepresentation of chimeric genes in regions associated with selective sweeps in comparison to duplicates is significant. Thus, chimeric genes are substantially more likely to be involved in selective sweeps than young duplicate genes and therefore offer a substantially richer source of genetic material for adaptation in the near term.

section*Discussion

Domain breaks and exon shuffling

Classic views on exon shuffling (Gilbert 1978, Patthy 2003) proposed that rearrangement of protein domains could result in peptides with novel functions, with introns providing easy boundaries for recombination and in-frame rearrangements. Furthermore, it is well established that different whole-domain combinations can result in functional novelty (Patthy

2003, Bashton and Chothia 2007, Vogel et al. 2004). However, in many cases we observe that chimera formation does not respect domain boundaries, and many of the breakpoints fall within exons, even when uniting portions of drastically different proteins. These mid-domain breaks have the potential to drastically interfere with protein function, and are often associated with human cancers (Kaye 2009, Mitelman, Johansson and Mertens 2007).

Still, mid-domain breaks can produce fully functional proteins when recombining distantly related paralogs (Mody, Weiner and Ramanathan 2009), often resulting in new phenotypes or catalytic abilities. However, the extent to which mid-domain breaks contribute to new, functional peptides is not fully established. Among our chimeric genes, those with mid-domain breaks do not create obviously dysfunctional peptides. *CG18853*, in spite of a mid-domain break, lies at the center of a selective sweep. Additionally, *Qtzl* has a segment that is inherited out of frame with respect to the parental gene, yet it was also involved in a selective sweep in *D. melanogaster* (Rogers et al. 2010).

While the locus of selection is difficult to determine for these sweeps, these results suggest that classical views on exon shuffling and protein splicing may need to be reexamined. Mid-domain breaks are often thought to be detrimental, resulting in dysfunctional peptides that can harm the cell. Yet, some of these peptides are found in the centers of strong selective sweeps, making it unlikely that they cause substantial cellular problems. While whole-domain recombinations can clearly produce new, functional peptides, these results suggest that mid-domain breaks could be equally important for development of novel functions both in an evolutionary context and in protein engineering.

Evolutionary novelty

Adaptation depends largely upon mutation. While single amino acid substitutions are extremely common, their ability to explore adaptive protein structures is quite limited (Carneiro and Hartl 2010). More unusual genetic combinations, while somewhat rare, are able to explore greater distances in protein folding space, rendering accessible more adaptive peaks (Bogarad and Deem 1999, Cui et al. 2002, Giver and Arnold 1998). Chimeric genes are well recognized for their ability to create unusual combinations of protein domains, and in some cases can even force mid-domain breaks that unite wholly unrelated proteins. While many of these combinations are likely to be dysfunctional (Voigt et al. 2002, Cui et al. 2002), some appear to be advantageous.

Beyond the combinations of active sites and binding domains, however, we have shown that chimeric genes offer rapid means of changing protein context. Expression profiles, cellular targeting, and trans-membrane orientation all offer different axes whereby chimeric genes can create new sequences or expression patterns with novel phenotypic effects. These contextual changes can in turn drive subsequent evolution of gene sequences. While not all such changes, or even most of such changes, might be beneficial, they should open up a great range of options that are available to natural selection. Hence, if extreme selective pressures require unusual genetic solutions, chimeric genes are likely to be an important source of extraordinary genes.

Theoretical models predict that initial adaptive steps come from mutations of large effect (Orr 2005), which are then followed by mutations of smaller effect that offer minor functional

adjustments. Chimeric genes provide for large-scale genetic change, which in some cases can translate to extreme selective effects. Estimates of selection coefficients for our chimeric genes can be as high as 1%, a massive selective impact for a population whose size is as large as that of *D. melanogaster*. The appearance of these genes is then often followed by a series of adaptive amino acid replacements that can modify the function of chimeric genes (Jones and Begun 2005, Jones, Custer and Begun 2005, Shih and Jones 2008). Up to 20% of chimeric genes that form may be selectively favored, and while somewhat less common than duplicate genes, chimeric genes form at high enough rates to provide a steady stream of adaptive changes (Rogers, Bedford and Hartl 2009). Therefore, we would suggest that chimeric genes, while seemingly unconventional, are important factors in adaptive evolution as well as serious contributors to genomic content.

Methods of formation

We have used rigorous definitions to identify chimeric genes which formed from two or more coding sequences in *D. melanogaster*. These requirements are not intended to identify all chimeric genes, but rather to provide a dataset of genes whose chimeric origins are well established. They do not capture sequences which recruit previously non-coding regions, chimeric transposable element constructs, or genes which recruit novel UTRs, which are known to form often in *C. elegans* and at at least modest rates in *D. melanogaster* (Katju, Farslow and Bergthorsson 2009, Zhou et al. 2008). As such, they may not capture the full diversity that can be achieved through chimeric gene formation. Yet, even this conservatively

defined dataset displays a strong case for genetic novelty and a considerable role in adaptive evolution.

All of the chimeric genes we observed in *D. melanogaster* appear to form through tandem duplications that have not respected gene boundaries (Rogers, Bedford and Hartl 2009), consistent with other work on chimeric genes in *D. melanogaster* (Zhou et al. 2008), but in stark contrast with the handful of chimeric retrogenes that have been identified in other species of *Drosophila* and a large number of retrogene chimeras in rice (Wang et al. 2006). This disparity is consistent with low levels of retroelement activity in the *D. melanogaster* lineage and is not influenced by our definition of chimeric genes. Other studies in mammalian genomes have identified chimeric genes that formed by ectopic recombination as well as retrogenes that have recruited novel exons (Wolf et al. 2009, Sedman et al. 2008, Opazo et al. 2009).

Retrogenes have been associated with novel expression in the testes and are often the targets of positive selection (Tracy et al. 2010, Quezada-Diaz et al. 2010, Betran, Bai and Motiwale 2006, Bai, Casola and Betran 2008). The absence of chimeric retrogenes from our dataset should make changes in gene expression less likely and could render selective sweeps less frequent. Still, we are able to identify a number of cases where DNA-level tandem duplications have resulted in expression changes, unusual peptide combinations, and selective sweeps, indicating that even these relatively unlikely candidates can be a valuable source of genomic changes.

Expression changes and pleiotropic constraints

Here, we have demonstrated that that chimeric gene creation can effect a number of different regulatory changes, including novel contexts for existing peptides and entirely distinct expression profiles for newly formed genes. As seen with *CG12592*, temporal expression patterns can change independently of expression profiles across tissues. Furthermore, through the addition or removal of cellular targeting signals, fine-scale localization of peptides can be modified or the membrane orientation can be reversed through the addition of hydrophobic domains. Many of these changes are associated with genes that have experienced selective sweeps. Thus, the formation of chimeric genes can quickly create broad changes in where a gene is expressed or targeted, likely influencing evolutionary outcomes.

While duplicate genes are known as key players in adaptive evolution and the origins of developmental complexity, they do not typically provide for a substantial source of adaptation in the near term. They do not commonly display qualitative expression changes or alter cellular targeting, and they are associated with selective sweeps far less often than chimeric genes. Duplicate genes create redundancy, which allows genes to diverge via neofunctionalization and subfunctionalization. However, each of these fates requires extremely long periods of time in which genes accumulate point mutations. Additionally, sequences that are expressed in multiple tissues or in both sexes may be highly constrained by pleiotropic effects (Van Dyken and Wade 2010, Yampolsky and Bouzinier 2010). New changes that can be advantageous in one tissue or life stage can cause detrimental effects in another context. Without the appropriate structural or regulatory changes, the peptide

may not be able to explore certain adaptive possibilities. If these mutations are rare, the duplicate gene may not acquire a novel function for a very long time and could be subject to decay before reaching a favored state. Hence, duplicate genes may not provide a timely source of adaptive material in the face of sudden or short-term selective pressures.

Chimeric genes, on the other hand, offer a means whereby single mutations can produce substantial flexibility that will allow organisms to explore a vast range of mutational space. The ability to change expression patterns along different axes immediately and independently can free chimeric genes from a large number of pleiotropic constraints and allow rapid evolution. Thus, chimeric genes can allow not only for more immediate changes, but they can also potentially allow sequences to explore a greater range of mutational space and adaptive possibilities.

Gene preservation

We have found signals of low diversity and highly skewed site frequency spectra surrounding three chimeric genes, *CG18853*, *CG81217*, and *Qtzl*, that are consistent with selective sweeps. One of these genes is associated with demonstrated phenotypic effects (Rogers et al. 2010) and all three display key expression changes. The parental genes appear to be distinct peptides in all three cases, and all contain widely different protein domains. This suggests that selection may commonly favor chimeric genes that form from drastically unrelated pepetides.

Examining the oldest chimeric genes, which have been preserved over long periods of time, we find that they commonly form from related proteins, but both parental genes and the

chimeric gene have distinct expression profiles. This pattern would suggest that expression changes could be essential to preservation of chimeric genes, consistent with theories of subfunctionalization and neofunctionalization of duplicate genes. Shuffling portions of distantly related proteins has been shown to produce novel phenotypic effects in yeast (Mody, Weiner and Ramanathan 2009), and it is possible that while these parental peptides appear similar, their chimeric rearrangements could produce fully distinct functions.

However, beyond the importance of gene expression changes, these results highlight the discordance between young, selectively favored chimeric genes, and the older, preserved chimeric genes. Such a disparity implies that in *D. melanogaster* the forces that shape genome content over long periods of time may differ from the forces that are active in short-term adaptation to newly arising selective pressures. While many new chimeric genes form from unrelated parental genes, virtually all of these types of chimeras seem to disappear over time, leaving only those that have formed from distant paralogs. This disparity is apparent even when adjusting chimeric gene search criteria to include chimeras that have subsequently duplicated (SI Text).

It has been proposed that genome content is largely dependent upon the likelihood with which new genes are able to fix in populations (Lynch 2007, Katju, Farslow and Bergthorsson 2009). These arguments rely on the assumption that genes once fixed will remain in the genome over very long time periods. A related and more common assumption is that neofunctionalization and adaptive subfunctionalization intrinsically confer preservation (reviewed in Hahn 2009, Innan and Kondrashov 2010). These altered genes, once fixed by

selection, will not readily be removed from the genome. Such assertions require that selective pressures on genes remain constant or that subsequent non-functionalizing mutations are rare.

Yet, the observed disparity between young and old chimeric genes in *D. melanogaster* implies that neither of these views can account for the number and type of chimeric genes that are preserved in the genome. Advantageous genes that confer novel functions may readily fix, but often will not be maintained. In the absence of selection to maintain newly fixed genes, a deletion-biased genome like *D. melanogaster* (Petrov, Lozovskaya and Hartl 1996) is likely to lose genetic factors that were once advantageous but are not currently favored. Moreover, fixation alone is not sufficient to result in the preservation of genes over long time periods. Rather, preservation occurs when long-standing selective pressures prevent the removal of new sequences. These instances can result either from partitioning ancestral gene functions or through the development of new functions. Hence, the distinction between genes that are maintained over time and those that are removed from the genome lies not solely in their functional differences relative to ancestral genes but more importantly relies heavily on the persistence of selective pressures.

Acknowledgements

We would like to thank Kalsang Namgyal and Ana M. Lyons for technical assistance and Michael B. Eisen, Trevor Bedford, Alexis S. Harrison, Sarah E. Irvin, Geoffrey F. Dilly, and Russ Corbett for helpful discussions. We also thank three anonymous reviewers for their

comments, which substantially improved the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology. 215:403–410.

Bai Y, Casola C, Betran E. 2008. Evolutionary origin of regulatory regions of retrogenes in *Drosophila*. BMC genomics. 9:241.

Bashton M, Chothia C. 2007. The generation of new protein functions by the combination of domains. Structure. 15:85–99.

Betran E, Bai Y, Motiwale M. 2006. Fast protein evolution and germ line expression of a *Drosophila* parental gene and its young retroposed paralog. Molecular Biology and Evolution. 23:2191–2202.

Bogarad LD, Deem MW. 1999. A hierarchical approach to protein molecular evolution. Proceedings of the National Academy of Sciences, USA. 96:2591–2595.

Carneiro M, Hartl DL. 2010. Colloquium papers: Adaptive landscapes and protein evolution. Proceedings of the National Academy of Sciences, USA. 107 Suppl 1:1747–1751.

Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. Proceedings of the National Academy of Sciences, USA. 99:809–814.

Adams et al. 2000. The genome sequence of *Drosophila melanogaster*. Science. 287:2185–2195.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology. 300:1005–1016.

Giansanti MG, Bucciarelli E, Bonaccorsi S, Gatti M. 2008. *Drosophila* SPD-2 is an essential centriole component required for PCM recruitment and astral-microtubule nucleation. Current Biology. 18:303–309.

Gilbert W. 1978. Why genes in pieces? Nature. 271:501.

Giver L, Arnold FH. 1998. Combinatorial protein design by *in vitro* recombination. Current Opinion in Chemical Biology. 2:335–338.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature. 445:82–85.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. The Journal of Heredity. 100:605–617.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland, Mass.: Sinauer Associates.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews. Genetics. 11:97–108.

Jones CD, Begun DJ. 2005. Parallel evolution of chimeric fusion genes. Proceedings of the National Academy of Sciences, USA. 102:11373–11378.

Jones CD, Custer AW, Begun DJ. 2005. Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics. 170:207–219.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics. 123:887–899.

Katju V, Farslow JC, Bergthorsson U. 2009. Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*. Genome Biology. 10:R75.

Kaye FJ. 2009. Mutation-associated fusion cancer genes in solid tumors. Molecular Cancer Therapeutics. 8:1399–1408.

Liu LX, Margottin F, Gall SL, Schwartz O, Selig L, Benarous R, Benichou S. 1997. Binding of HIV-1 Nef to a novel thioesterase enzyme correlates with Nef-mediated CD4 down-regulation. The Journal of Biological Chemistry. 272:13779–13785.

Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science. 260:91–95.

Lynch M. 2007. The Origins of Genome Architecture. Sunderland, Mass.: Sinauer Associates.

McLennan AG. 1999. The MutT motif family of nucleotide phosphohydrolases in man and human pathogens (review). International Journal of Molecular Medicine. 4:79–89.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer. 7:233–245.

Mody A, Weiner J, Ramanathan S. 2009. Modularity of MAP kinases allows deformation of their signalling pathways. Nature Cell Biology. 11:484–491.

Ohshima K, Igarashi K. 2010. Inference for the initial stage of domain shuffling: tracing the evolutionary fate of the *PIPSL* retrogene in hominoids. Molecular Biology and Evolution. 27:2522–2533.

Opazo JC, Sloan AM, Campbell KL, Storz JF. 2009. Origin and ascendancy of a chimeric fusion gene: the beta/delta-globin gene of paenungulate mammals. Molecular Biology and Evolution. 26:1469–1478.

Orr HA. 2005. The genetic theory of adaptation: a brief history. Nature Reviews Genetics. 6:119–127.

Patthy L. 1999. Genome evolution and the evolution of exon-shuffling–a review. Gene. 238:103–114.

Patthy L. 2003. Modular assembly of genes and the evolution of new functions. Genetica. 118:217–231.

Peisajovich SG, Garbarino JE, Wei P, Lim WA. 2010. Rapid diversification of cell signaling phenotypes by modular domain recombination. Science. 328:368–372.

Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature. 384:346–349.

Qiu J, Hardin PE. 1995. Temporal and spatial expression of an adult cuticle protein gene from *Drosophila* suggests that its protein product may impart some specialized cuticle function. Developmental Biology. 167:416–425.

Quezada-Diaz JE, Muliyil T, Rio J, Betran E. 2010. *Drcd-1 related*: a positively selected spermatogenesis retrogene in drosophila. Genetica. 138:925–937.

Rogers RL, Bedford T, Hartl DL. 2009. Formation and longevity of chimeric and duplicate genes in *Drosophila melanogaster*. Genetics. 181:313–322.

Rogers RL, Bedford T, Lyons AM, Hartl DL. 2010. Adaptive impact of the chimeric gene *Quetzalcoatl* in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences, USA. 107:10943–10948.

Sedman L, Padhukasahasram B, Kelgo P, Laan M. 2008. Complex signatures of locus-specific selective pressures and gene conversion on human growth hormone/chorionic somatomammotropin genes. Human Mutation. 29:1181–1193.

Shih HJ, Jones CD. 2008. Patterns of amino acid evolution in the *Drosophila ananassae* chimeric gene, *siren*, parallel those of other *Adh*-derived chimeras. Genetics. 180:1261–1263.

Singh ND, Arndt PF, Petrov DA. 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics. 169:709–722.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Tatusova TA, Madden TL. 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters. 174:247–250.

Thornton KR. 2007. The neutral coalescent process for recent gene duplications and copy-number variants. Genetics. 177:987–1000.

Tracy C, Rio J, Motiwale M, Christensen SM, Betran E. 2010. Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in *Drosophila*. Genetics. 184:1067–1076.

Tusnady GE, Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. Journal of Molecular Biology. 283:489–506.

Tusnady GE, Simon I. 2001. The HMMTop transmembrane topology prediction server. Bioinformatics. 17:849–850.

Van Dyken JD, Wade MJ. 2010. The genetic signature of conditional expression. Genetics. 184:557–570.

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multidomain proteins. Current Opinion in Structural Biology. 14:208–216.

Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. 2002. Protein building blocks preserved by recombination. Nature Structural Biology. 9:553–558.

Wang W, Zheng H, Fan C, et al. (14 co-authors). 2006. High rate of chimeric gene origination by retroposition in plant genomes. The Plant Cell. 18:1791–1802.

Wolf A, Millar DS, Caliebe A, et al. (18 co-authors). 2009. A gene conversion hotspot in the human growth hormone (GH1) gene promoter. Human Mutation. 30:239–247.

Yampolsky LY, Bouzinier MA. 2010. Evolutionary patterns of amino acid substitutions in 12 *Drosophila* genomes. BMC Genomics. 11 Suppl 4:S10.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. Genome Research. 18:1446–1455.

Table 1: Expression Pattern of Chimeric Genes

Chimera	5′ Parental	3′ Parental	Defines Expression Pattern	d_S
CG31904	*CG13796*	*CG7216*	like 5′ end	0.003
CG18853	*CG12822*	*CG11205*	like 5′ end	0.004
CG32318	*CG9191*	*CG9187*	like 5′ end	0.008
CG31864	*CG12264*	*CG5202*	Distinct	0.010
CG12592	*CG18545*	*CG12819*	Distinct	0.017
CG31687	*CG2508*	*CG31687*	Distinct	0.025
CG18217	*CG17286*	*CG4098*	like 5′ end	0.027
CG30457	*CG10953*	*CG13705*	Distinct	0.125
CG17196	*CG17197*	*CG17195*	like 3′ end	0.414
CG11961	*CG9416*	*CG30049*	Distinct	0.501
CG3978	*CG9656*	*CG10278*	Distinct	0.576
CG6844	*CG5610*	*CG11348*	Distinct	0.727
CG6653	*CG31002*	*CG17200*	like 5′ end	0.743
CG31688	*CG33124*	*CG8451*	like 5′ end	0.513

Table 2: Expression Patterns of *CG12592* and Parental Genes

	CG18545	*CG12592*	*CG12819*
Male Head	-	-	-
Testes	+	+	+
Male Carcass	-	-	+
Embryos 0-14hr	-	+	+
Embryos 14-24hr	-	-	-
early larvae	-	+	+
late larvae	-	-	-
Pre-pupae	+	+	+
late pupae	+	+	+
adult males	+	+	+
adult females	-	+	+

+ Present, - Absent

Table 3: Tajima's D for Chimeric and Parental Genes

Chimera	Tajima's D	5′ Parental	Tajima's D	3′ Parental	Tajima's D
CG31904	-1.15	*CG13796*	-0.91	*CG7216*	-0.47
CG18853	-1.90	*CG12822*	-2.05	*CG11205*	-2.07
CG32318	-0.58	*CG9191*	0.13	*CG9187*	-0.46
CG31864	-2.49	*CG12264*	-2.22	*CG5202*	-2.01
CG12592	-2.01	*CG18545*	-2.04	*CG12819*	-1.96
CG31687	-1.50	*CG2508*	-0.75	*CG31687*	-1.50
CG18217	-2.28	*CG17286*	-2.37	*CG4098*	-2.54
CG30457	-0.98	*CG10953*	-1.18	*CG13705*	-0.90
CG17196	-1.24	*CG17197*	-1.31	*CG17195*	-1.30
CG11961	-0.82	*CG9416*	-0.30	*CG30049*	-0.25
CG3978	-0.91	*CG9656*	-1.35	*CG10278*	-1.62
CG6844	-1.09	*CG5610*	-0.98	*CG11348*	-1.02
CG6653	0.07	*CG31002*	-0.50	*CG17200*	-0.23
CG31688	-1.08	*CG33124*	-1.39	*CG8451*	-0.82

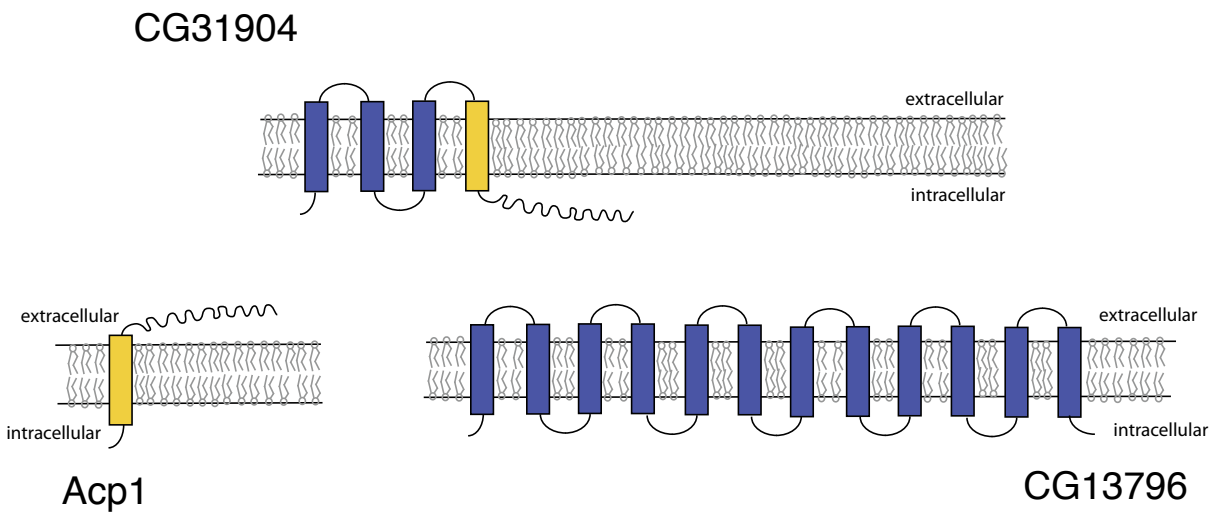

Figure 1: Membrane-bound domains for CG31904. CG31904 combines three trans-membrane helices from the predicted neurotransmitter transporter CG13796 and a single trans-membrane helix from Adult cuticular protein 1 (Acp1). The resulting peptide carries the long hydrocarbon chain from Acp1 which now faces inside rather than outside the cell

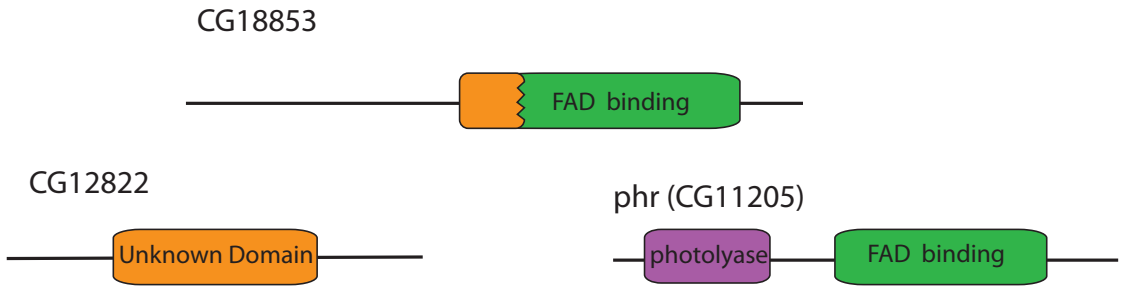

Figure 2: Mid-domain breaks in CG18853. Full length of the peptide is shown with a black line. Conserved domains are depicted with shaded rectangles. CG18853 formed during a tandem duplication event that did not respect boundaries of genes or conserved protein domains. The parental peptide CG12822 carries a conserved domain of unknown function that is found in vertebrates as well as in multiple bacteria. The human ortholog of CG12822, Nef-associated protein 1, is a thioesterase that interacts with HIV protein Nef. The formation of CG18853 combined a portion of this domain with an FAD-binding domain to produce a new peptide that lies at the center of a selective sweep in *D. melanogaster*.

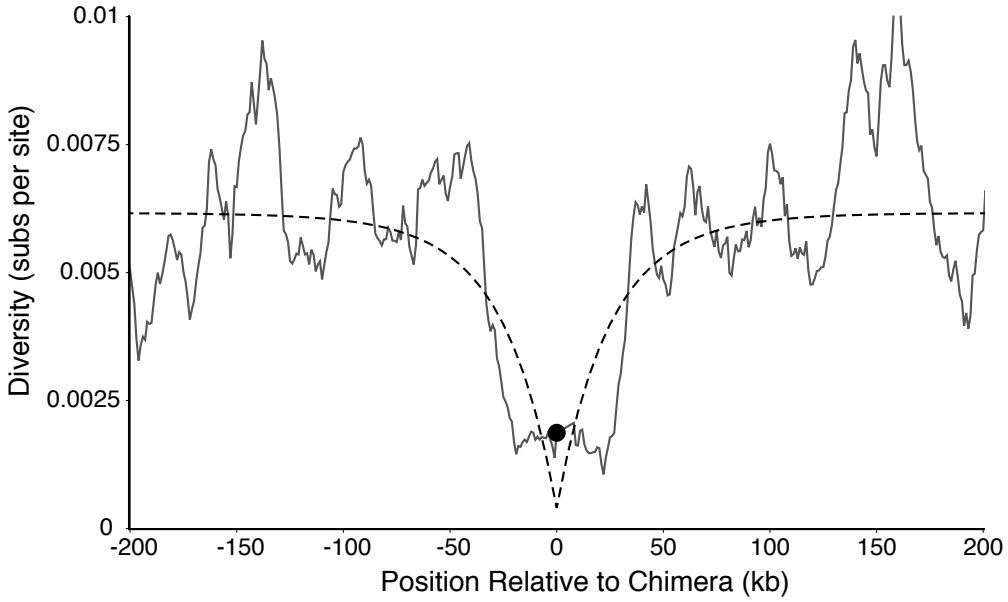

Figure 3: Local diversity π, measured as substitutions per site, surrounding *CG18217* (solid line) fitted with the expectation after a selective sweep (dashed line). The fitted curve describes a selective sweep with $s = 0.006$ which occurred 20,000 years ago. The reduction in diversity spans 40 kb, which includes multiple gene sequences. The chimeric gene CG18217 lies at the center of the selective sweep.

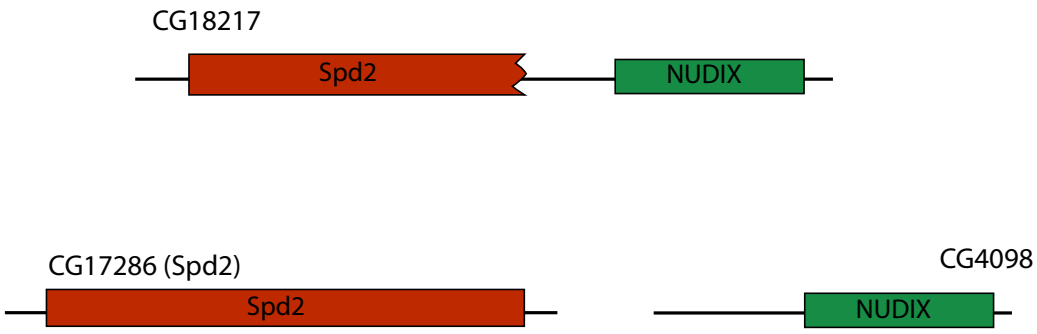

Figure 4: Domain structure for CG18217. Full length of the peptide is shown with a black line. Conserved domains are depicted with shaded rectangles. The chimeric gene *CG18217* was created when a tandem duplication united a portion of a DNA-repair gene *CG4098* with a portion of the spindle-formation gene *Spd-2*. Chimeric gene formation has disrupted the Spd-2 protein and combined it with a NUDIX DNA-repair domain. The new 5′ end of the gene now confers expression in a greater number of tissues.

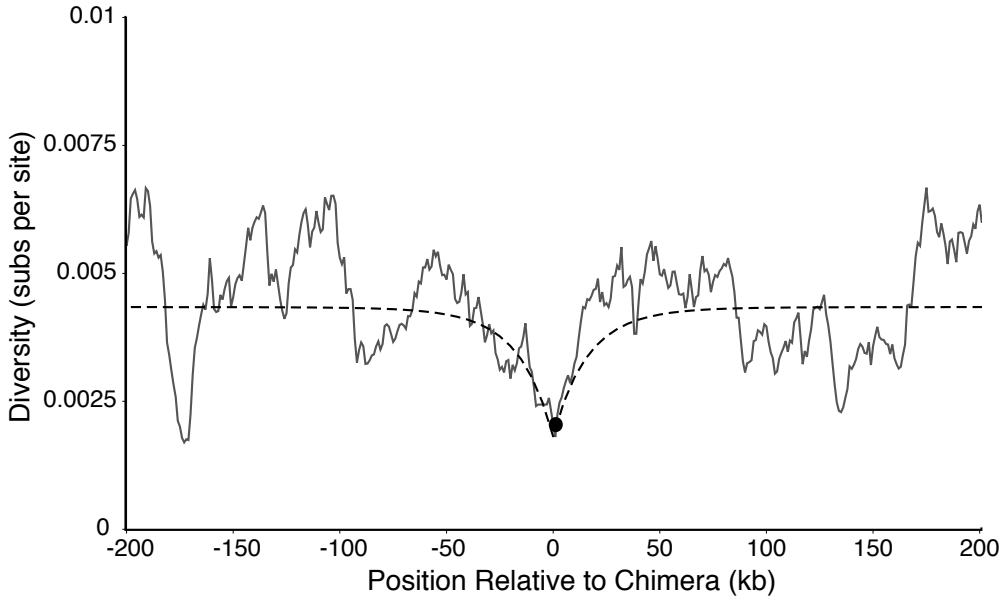

Figure 5: Local diversity π, measured as substitutions per site, surrounding *CG18853* (solid line) fitted with the expectation after a selective sweep (dashed line). The fitted curve describes a selective sweep with $s = 0.0025$ which occurred 200,000 years ago. The reduction in diversity spans 45 kb, which includes multiple genes. The chimeric gene CG18853 lies at the center of the selective sweep.